Araujo, María Caridad; Rubio-Codina, Marta; Schady, Norbert Rüdiger

Working Paper

70 to 700 to 70,000: Lessons from the Jamaica experiment

IDB Working Paper Series, No. IDB-WP-1230

Provided in Cooperation with:
Inter-American Development Bank (IDB), Washington, DC

Suggested Citation: Araujo, María Caridad; Rubio-Codina, Marta; Schady, Norbert Rüdiger (2021) : 70 to 700 to 70,000: Lessons from the Jamaica experiment, IDB Working Paper Series, No. IDB-WP-1230, Inter-American Development Bank (IDB), Washington, DC, https://doi.org/10.18235/0003210

This Version is available at:
https://hdl.handle.net/10419/237505

IDB WORKING PAPER SERIES Nº IDB-WP-1230

70 to 700 to 70,000: Lessons from the Jamaica Experiment

M. Caridad Araujo
Marta Rubio-Codina
Norbert Schady

Inter-American Development Bank
Social Protection and Health Division

April 2021

70 to 700 to 70,000: Lessons from the Jamaica Experiment

M. Caridad Araujo
Marta Rubio-Codina
Norbert Schady

April 2021

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Araujo, Maria Caridad.
70 to 700 to 70,000: lessons from the Jamaica experiment / M. Caridad Araujo, Marta Rubio-Codina, Norbert Schady.
p. cm. — (IDB Working Paper Series ; 1230)
Includes bibliographic references.
1. Home-based family services-Jamaica. 2. Home-based family services-Colombia. 3. Home-based family services-Peru. 4. Child development-Government policy-Jamaica. 5. Child development-Government policy-Colombia. 6. Child development-Government policy-Peru. I. Rubio-Codina, Marta. II. Schady, Norbert Rüdiger, 1967- III. Inter-American Development Bank. Social Protection and Health Division. IV. Title. V. Series.
IDB-WP-1230

http://www.iadb.org

Scl-sph@iadb.org
www.iadb.org/SocialProtection

70 to 700 to 70,000: Lessons from the Jamaica Experiment

M. Caridad Araujo
Marta Rubio-Codina
Norbert Schady

2021

70 to 700 to 70,000: Lessons from the Jamaica Experiment

M. Caridad Araujo Marta Rubio-Codina Norbert Schady

Abstract[†]

This document compares three versions of the same home visiting model, the well-known Jamaica model, which was gradually scaled-up from an efficacy trial ('proof of concept') in Jamaica, to a pilot in Colombia, to an at-scale program in Peru. It first describes the design, implementation and impacts of these three programs. Then, it analyzes the threats to scalability in each of these experiences and discusses how they could have affected program outcomes, with a focus on three of the elements of the economic model of scaling in Al-Ubaydli, et al. (Forthcoming): appropriate statistical inference, properties of the population, and properties of the situation. The document reflects on the lessons learned to mitigate the threats to scalability and on how research and evaluation can be better aligned to facilitate and support the scaling-up process of early child development interventions. It points out those attributes that interventions must maintain to ensure effectiveness at scale. Similarly, political support is also identified as indispensable.

Key Words: home visiting, quality, child development, monitoring, scalability.

JEL codes: I38, J24, O15

[†] We thank Julieth Parra for outstanding research assistance in the preparation of this document.

Contents

1. Introduction

This document describes three efforts to implement and evaluate variants of the same home visiting intervention, aimed at improving parent-child interactions and child development, in Jamaica, Colombia, and Peru. The document is descriptive. While each of these efforts was evaluated independently by randomized trials, any statistical comparison of program characteristics and impacts across them is difficult with the data available. Rather, the goal of our report is to document the process whereby the intervention was gradually scaled-up, from an efficacy trial in Jamaica, to a pilot designed to allow replicability at scale in Colombia, to an at-scale government program in Peru.[1]

The first of these programs was implemented in Kingston, Jamaica, during the mid-1980s. A research team developed a set of play materials and activities to improve developmental outcomes for malnourished children, through psychosocial stimulation and play. It was delivered by paraprofessional community health workers, who, during weekly home visits to sixty-four children, modelled play activities and responsive adult-child interactions. Findings from a randomized control trial (RCT) showed large effects (0.88 standard deviations [SDs] on overall development) after 24 months of visits (Grantham-McGregor et al., 1991).[2] To date, those children have been followed for over 30 years. Twenty years after the program ended, those who had been randomly assigned to treatment had completed more schooling, had higher test scores, were more likely to be employed, and had wages that were 25% higher than those assigned to the control group (Gertler et al., 2014). They also had improved cognitive, personal-social, and mental health outcomes (Walker et al., 2011).

The second program we describe is the adaptation of the Jamaican model implemented as a pilot in Colombia in 2010-2011. It primarily aimed to test whether the home visiting intervention could be integrated into the infrastructure of a national social protection program as a pathway toward scalability. Home visits were delivered by local leaders, known as *madres líderes* (leader mothers), who had been elected by their communities to be liaisons with the administrators of Colombia's nationwide conditional cash transfer (CCT) program. A second goal of the evaluation was to identify the mechanisms underlying impacts on child development.

About 720 children received home visits for 18 months. A cluster randomized control trial (CRCT) found effects on child development in the short run: 0.26 SDs on cognition, 0.22 SDs on receptive language, and 0.18 SDs on an aggregate index combining cognition, receptive and expressive language, and fine motor development (Attanasio et al., 2014). Statistical modelling showed that effects were mainly driven by increases in parental material and time investments in their children—namely, the number of play materials and the variety of play activities the child engaged

[1] The intervention was developed and first implemented over several trials in Jamaica, including the one we describe at length; and later adapted, improved, and replicated, on 18 occasions in 14 different countries to date. All these replications have evaluations attached. This process led to the development of the *Reach Up and Learn* package (Grantham-McGregor & Smith, 2016), which includes improved materials for the adaptation, implementation, and training of the model in low-resource settings (Walker et al., 2018). We focus on the experiences in Jamaica, Colombia, and Peru since they represent landmarks in the process of scaling up the model.

[2] Overall development included the following areas: the locomotor (gross motor), hand and eye coordination (fine motor), hearing and speech (language), and performance (cognition) subscales of the Griffiths Mental Development scales.

in with an adult (Attanasio et al., 2020). However, there was no evidence that effects persisted two years after the program ended (Andrew et al., 2018).

The third program we discuss, the Cuna Más program in Peru, was modelled on Colombia's program.[3] It was truly an intervention at scale, implemented by the Ministry of Social Inclusion and Development. Cuna Más aimed to reach all children 0-3 years of age in rural localities with high rates of poverty and chronic malnutrition, and was geographically targeted to areas where the nationwide CCT program was operating. Over the period of its evaluation, Cuna Más extended coverage from 5,338 children in April 2013 to 67,332 children in December 2015. It is currently serving over 115,000 children (Figure 1).

The evaluation, by CRCT, included 3,530 children in the treatment group and showed modest but significant improvements (0.10 SDs), on an aggregate index of child development that combined cognition, language, fine motor, gross motor, and personal-social abilities (Araujo et al., 2019). Impacts were larger (0.14 SDs) on children assigned to the program who received at least one visit.[4] Caregivers in the treatment group were also found to play with their children more often and engage less in punitive discipline strategies than those in the control group. In addition, among children 36 months and older, the age at which they are no longer eligible for visits, children randomly assigned to treatment were 9 percentage points more likely to be enrolled in preschool, relative to enrollment rates of 50% in the control group.

While the core of the intervention was common across countries, their design and implementation differed in several ways. The Jamaica project was an efficacy trial, highly controlled. The experience in Colombia sought to adapt and implement the model in a manner that could be replicated at scale, relying on existing community resources, and with low-intensity supervision. Both experiences were carried out in the context of research projects. Peru, in turn, was an at-scale implementation of the model by a government agency, with a very rapid expansion in coverage.

We frame our comparison of the three programs using elements of the model of scaling presented in Al-Ubaydli, et al. (Forthcoming): (a) appropriate statistical inference; (b) properties of the population; and (c) properties of the situation.[5] We organize our discussion of the design, implementation, and impacts of the Jamaica, Colombia, and Peru programs around these elements.

The rest of the document is organized as follows. The first section describes the three interventions. The next compares them using insights from the Al-Ubaydli et al. (Forthcoming) model to discuss the large declines in program effect sizes as the intervention was scaled up. The final section reflects on lessons learned.

[3] In addition to the home visiting program, *Cuna Más* provides daycare services in low-income urban areas. For simplicity, in this document we refer to the home visiting program as *Cuna Más*, without making explicit that we are only referring to one of the two services the program provides.
[4] A third of the intended population in the evaluation sample did not receive home visits. The reasons are discussed later in this document.
[5] The model presents a fourth element, spillovers and general equilibrium effects, not analyzed in this document.

2. The Jamaica, Colombia, and Peru Experiences

We begin with a more detailed description of the main program design features of the three home visiting interventions under study, presented in Table 1.

All three shared the same (a) mode of implementation—weekly visits delivered by local paraprofessionals with low average education levels; (b) methodology—play-based visits that sought to promote child development and caregiver-child interactions; and (c) a core curriculum of structured activities and play materials—mostly low-cost home-made toys, blocks, picture books, puzzles, sorting and matching activities—with adaptations to the local context.

Activities in the curriculum were arranged in order of difficulty to facilitate scaffolding (i.e., the practice of teaching a child something new by applying things they already know) and were organized around the family's daily routines. The aim was to work through caregivers, by building a positive relationship with them, in order to strengthen their skills and enjoyment in promoting child development and helping their children learn (Walker et al., 2018). To this end, the home visitor (HV) demonstrated activities that introduced new concepts and challenges during the visit.

Specifically, the HV modelled actions to encourage caregivers to engage in play and conversation with their child and to respond to their vocalizations and actions. They also encouraged praising the child and celebrating the child's efforts and achievements in order to promote their self-esteem and socio-emotional development. Special emphasis was placed on listening to the caregiver, seeking their opinion, and giving them encouragement. Ultimately, the program sought to generate behavioral changes in childrearing practices that lasted beyond the life of the intervention.[6]

Despite these similarities, there were important implementation differences across programs. The Jamaican program enrolled children at ages 9-24 months and lasted for 24 months, therefore serving some children until they were 4 years of age. In Colombia, children received visits for a shorter period—18 months, starting at ages 12-24 months up until they were 30-42 months old. The Peruvian program was the most ambitious in terms of intended duration and starting age, enrolling pregnant mothers and children from birth to 24 months-of-age, with visits beginning in pregnancy and running until the child turned 3 years old.

The target population also differed across these interventions. On average, children who received the intervention in Jamaica and Peru were living in higher-poverty communities—in urban slums in Jamaica and in poor remote rural communities in Peru—and experiencing higher rates of chronic malnutrition[7] than children who received the intervention in Colombia. This may matter in terms of the potential of the intervention to impact developmental outcomes, relative to baseline

[6] While caregivers are critical for the intervention's success—as are improving their self-esteem and self-efficacy—the model does not specifically target caregiver health or economic wellbeing, nor does it provide direct caregiving. In this regard, the programs are different from the Nurse Family Partnership program (Olds, 2010) and Early Head Start (Love et al., 2013), two of the most researched home visiting models in the US.

[7] This is children who had height-for-age more than two SDs below the median height-for-age in the WHO reference population of well-nourished children.

levels and the quantity and quality of parental investments in children. Earlier studies have shown that, in early childhood, the impact of interventions is often larger amongst children of lower socioeconomic status (Bitler et al., 2014 in the US; and Havnes & Mogstad, 2015 in Norway), although that is not always the case (Nores et al., 2019 in Colombia).

Table 2 presents information on *program implementation*, mainly from administrative records. Scale is orders of magnitude apart across programs, which results in important differences.

Some of the implementation variables, such as dosage or HV turnover, can be thought of as proxies of structural quality. Unlike Jamaica and Colombia, Peru faced substantial challenges in reaching all families and encouraging take-up (66% of children in the evaluation treatment group received at least one visit), guaranteeing dosage (half of the planned visits took place), and retaining HVs and supervisors (who on average stayed on the job for 17-18 months). Such challenges were possibly due to the large number of program beneficiaries, who were also spread over a vast geographical area.

The three programs were evaluated by means of RCTs, which are compared in Table 3. In Jamaica and Colombia, the evaluations were designed and led by researchers. The case of Peru was somewhat different. The ministry in which Cuna Más was housed, the Ministry of Social Development and Inclusion, had a strong technical mandate to evaluate its interventions that was closely aligned with a results-based budgeting process led by the Ministry of Finance. The decree whereby Cuna Más was created required the program had to have an experimental impact evaluation, and only upon its completion the program could become a permanent line item in the government budget.

The Jamaican and Colombian evaluations used full diagnostic tests to assess program outcomes—the Griffiths Mental Development Scales in Jamaica and the Bayley Scales for Infant and Toddler Development in Colombia. These were longer and more complex to apply—they had more items and had to be administered by professionals. They were also more costly to administer, requiring the payment of per-use license fees and the purchase of expensive kits of materials. In Peru, on the other hand, a screener test, the Ages and Stages Questionnaires, was used because administration of full diagnostic assessments was deemed too complex and expensive—especially given the much larger sample of the evaluation. Screeners are designed to identify children with developmental delays and thus focus on the lower tail of the distribution of abilities, which may reduce the capacity of the test to detect impacts.[8] Moreover, in Peru some of the items were administered by caregiver report, which could introduce biases as caregivers in the treatment group might be (a) more knowledgeable about child development and more aware

[8] The comparability of effect sizes across studies is limited not only because they each use different instruments to measure outcomes, but also because these outcomes are scaled with the SD of each sample. If the same outcome measure had been used in all three evaluations but such a measure did not have a comparable norm or scale, then any comparison of effect sizes would still be limited because of differences in the distribution of developmental outcomes across populations (and thus variances). If the same measure had been used and if this measure had a norming sample, even if it was from an external population, relative comparisons would be cleaner, although they could still be affected by differences between the study and norming sample distributions. The use of a common measure, globally normed and culturally neutral, such as the long form of the Global Scales of Early Development (GSED; Cavallera et al., 2019), would allow for meaningful comparisons of effects sizes across evaluations. However, development of the GSED or similar global measures for programmatic evaluation is still in progress.

of the child's achievements (*observation bias*) and/or (b) more inclined to make optimistic claims on these achievements so as to report on intervention success (*desirability bias*). The Peru and Colombia evaluations had similar levels of attrition between baseline and endline measurements.

3. Identifying Threats to Scalability

This section uses the economic model of scaling in Al-Ubaydli et al. (Forthcoming), which in turn is based on Al-Ubaydli et al., 2019, to analyze the Jamaica parenting intervention and subsequent efforts to adapt and bring it to scale in Colombia and Peru. It discusses whether the threats to scalability proposed in the model were present in each of these experiences and reflects on how they could have affected outcomes. It does so, first, by considering differences between Jamaica and Colombia ("from 70 to 700"); this discussion also draws on the implementation and evaluation of other pilots in Bangladesh, India, and China. We then turn to differences between Colombia and Peru ("from 700 to 70,000").

3.1 From 70 to 700: Jamaica and Colombia

As Table 3 shows, the short-term effects of the efficacy trial in Jamaica (0.88 SDs) are substantially larger in magnitude than those estimated for the pilot in Colombia (0.18 SDs). This implies that, in moving "from 70 to 700," the impact fell by about four-fifths. What could explain such a dramatic difference?

Problems of Statistical Inference

One possibility discussed in the Al-Ubaydli et al. model is problems of *statistical inference*. Concerns with statistical inference could arise if the Jamaica trial simply reflected a "lucky draw" among all possible draws from a given population—specifically, a draw from the right tail of the distribution of possible impacts. However, this does not appear to have been the case. In addition to the best-known of the Jamaican trials, which had an effect size of 0.88 SDs, there were at least three other replications with an RCT in Jamaica, all listed in Table 4. Effect sizes in two of these trials were larger (0.94 and 1.18 SDs, respectively) and only in one of them was the effect much smaller (0.28 SDs) and not significant. Seen in this light, there does not appear to be anything unusual in the results of the "Jamaica trial."[9]

That said, however, there *is* something unusual about the results from Jamaica—not just relative to those found in Colombia, but also to results from replications in (a) Bangladesh, where three efficacy trials have been carried out with effect sizes between 0.25 and 0.38 SDs on the Mental Development Index of the Bayley Scales (second edition); (b) India, where a recent replication trial had an impact of 0.21 SDs on an index combining cognition, receptive and expressive language, and fine motor development as assessed on the Bayley Scales (third edition); and (c) China, where a pilot intervention loosely based on the Jamaican model found effects of 0.17 SDs

[9] We put the term in quotation marks because the best-known of the Jamaica trials is not the only trial of the intervention in Jamaica, nor was it even the first one (so we cannot describe it as the "original" Jamaica trial). It is simply the best-known—largely, because of the long-term follow-up.

on an index that combined cognition, receptive and expressive language, fine motor development, gross motor, and socio-emotional development from the Bayley Scales.[10]

This can be seen clearly in Table 4 and Figure 2. The figure graphs the average effect size found in a given evaluation (on the vertical axis) with the (log of) the population covered by the program (on the horizontal axis). The color of the circles varies by country, and an empty circle indicates that a given impact is not significant statistically. As shown, the average effect sizes fall as the number of children covered by the intervention increases. Relatedly, the effects found in Jamaica (earliest evaluations), although variable, are on average substantially larger than those found elsewhere.

Figure 2 thus raises several possibilities that could account for the decline in effect sizes as the Jamaica program was gradually brought to scale. One is a particular problem of statistical inference: the instrument used in the Jamaica evaluations—or some items within it—may have been more sensitive to the kinds of changes in child development brought about by visits. Diagnostic developmental tests, such as the Griffiths or the Bayley Scales, share similar items since they all aim to measure the same construct: overall child development. However, the relative emphasis on each developmental domain, as well as the specific items included and their phrasing, do vary from test to test. Future analysis that carefully studies item-level performance scores in each evaluation would be useful to understand whether this can explain, in part, differences in the effect sizes between Jamaica and elsewhere.

Representativeness of the Population

Another issue raised in Al-Ubaydli et al. (2019) is related to the *representativeness of the population* and, more generally, whether the project participants were similar across interventions. In the Jamaica trial, participants were drawn from a census that identified children ages 9-24 months who were chronically malnourished, had birthweight above 1.8 kg, did not have twin siblings, did not have a detectable disability, had mothers with low levels of education, and lived in poor housing conditions in low-income neighborhoods of Kingston. It is possible that the fact that children in Jamaica were on average poorer than those in Colombia partly explains the smaller effect sizes found in Colombia. However, families who benefitted from the intervention in Bangladesh and India were, if anything, even more disadvantaged than those in Jamaica, so this does not appear to be a sufficient explanation for the smaller effect sizes that were observed as the Jamaica model was applied elsewhere.

Another explanation, perhaps, is a possible selection of the population on expected gains. Later evaluations—in particular, those carried out in Colombia and Peru—carefully discuss and quantify the number of families who were offered home visits but turned them down. The reported estimates are Intent-to-Treat (ITT), which means that they include all families randomized into the treatment group regardless of whether they were treated in practice. This is not the case in the Jamaica trials. None of the Jamaica studies report whether some families turned down visits and,

[10] For discussion of a program in China that more closely mirrored the Reach Up model, see Heckman, J., Liu, B. & Zhou, J. (Forthcoming).

if so, whether replacement families were selected in some way. It seems likely that, in all three countries, families who expected to benefit most from the program would be more likely to participate. If some families in Jamaica did turn down visits, selection on gains could be part of the explanation for the larger effects found there than elsewhere, since these families were not included in the evaluation results.

Representativeness of the Situation

We turn next to a discussion of the *representativeness of the situation.* The Jamaica project was not meant to be representative of a real-world situation, but rather to be a proof of concept. The Colombia pilot was an attempt to deliver visits in a scalable manner—this is, in a manner that could be replicated at scale. Several challenges during implementation could have compromised fidelity when scaling up the Jamaican model in Colombia.

Data presented in Tables 1 and 2 shows differences in the nature and quality of the supervision received by HVs. While the Jamaica trial had only three HVs and two supervisors, the Colombian pilot involved 144 HVs (164, accounting for turnover), and six supervisors. Not only were the numbers of staff involved larger in Colombia, but so were the supervision ratios. The HVs-to-supervisor ratio was 1.5 in Jamaica, compared to 24 in Colombia. Moreover, supervisors offered weekly in-person mentoring sessions in Jamaica, compared to every 7-10 weeks in Colombia. In Jamaica, the researchers were also supervisors, so they had substantially better qualifications and expertise than the supervisors in Colombia, who were university graduates with a diversity of backgrounds. The direct involvement in implementation and supervision of the researchers in Jamaica may also be part of the explanation. The researchers involved had developed the home visiting curriculum and methodology, had carefully thought about how to ensure that it was appropriate in the setting where it was implemented, and knew exactly what the home visits should look like. All of these conditions are likely to have been present to a smaller degree in other replications—including in Bangladesh, Colombia, India and China—than was the case in the “Jamaica trial”.

In sum, there are a number of possible explanations for the large decline in effect sizes in moving “from 70 to 700.” First, there were differences in the instruments used to measure child development, and some of the measured differences in effect sizes may simply be a statistical artifact. Second, there may be differences in the population that received home visits across settings. We do not believe that differences in socioeconomic status or child health are likely to be the main reason for the larger effect sizes in Jamaica than elsewhere since beneficiary children in other study settings, such as Bangladesh or India, were of equally poor or poorer backgrounds. Possible (undocumented) selection on expected benefits, perhaps more so in Jamaica than elsewhere, could be a part of the explanation. More generally, it could be that there is something inherent to the model, its design, or its form of delivery, that is more appropriate to Jamaica (or to the gaps of children in Jamaica), and that this explains some of the differences in effect sizes we observe. Third, there were underlying differences in how the program was implemented in different settings. Some of this is perhaps unsurprising—the Jamaica evaluations were ‘proof of concept’ trials, while the Colombia intervention sought to implement and evaluate an intervention

that was at least in principle *scalable*. The ratio of supervisors per HVs was much higher in Jamaica than in Colombia, as was the frequency of in-person supervision, and the involvement of the researchers in the implementation of the intervention.

3.2 From 700 to 70,000: Colombia and Peru

We next compare the Colombia and Peru programs, highlighting whether *additional* threats to scalability may have been present during the rapid scaling-up of the model.

The Cuna Más program in Peru was an at-scale real-world intervention run by a newly created Ministry. During the 2011 election campaign, candidate Ollanta Humala announced that, if he were elected, he would start a program for young children. After his election, Humala created a new Ministry with a new flagship program, Cuna Más. This Ministry was staffed with highly technical people who actively drew on international expertise (including that of the authors of this document and of the researchers involved in the Jamaican and Colombian evaluations) to design a program that was appropriate for rural areas in Peru. Authorities had to deliver results within five years, the length of the presidential term.

Home visiting programs are intensive in personnel, and staff selection and ratios are a key consideration for quality, costs, and scalability. In Peru, there were not enough people in the participating communities that met the technical qualifications required to be a HV, as originally specified. As a result, Cuna Más had to adapt these requirements, which meant hiring HVs with lower education levels than initially desired. This may have negatively affected quality, and it certainly delayed the process of adapting guidelines and of selecting HVs. Similarly, not all supervisors in Peru had tertiary education, as initially required and as was the case in Colombia.

Importantly, despite its scale, the Peruvian program maintained relatively low ratios of families per HV—10, on average—and HVs per supervisor—also 10, on average. This was a deliberate decision taken by senior staff at Cuna Más, in an effort to shield quality and to operate in disperse, poorly connected localities. Both ratios were substantially lower than those in Colombia.

Scaling up home visiting programs also requires rethinking the delivery of pre- and in-service training. Pre-service training was much shorter in Peru than in Colombia: In Colombia, HVs received two weeks of pre-service training, including practice visits, plus an additional week of training one-to-two months after the program had started; in Peru, on the other hand, HVs were given four days of pre-service training, without any practice visits. Turning to supervisors, in Colombia they received six weeks of training, compared to only nine days in Peru. Moreover, there may have been a greater loss of fidelity in training in Peru given the much larger number of people who had to be trained (and retrained due to frequent turnover) in a shorter amount of time.[11]

[11] In both countries, the training followed a cascade model, where each level trained the next level down the implementation ladder. In Colombia, two psychologists, who also supported adaptation of the program, trained the 6 supervisors, who in turn trained the

In both Colombia and Peru (as well as in Jamaica), in-service training for HVs took the form of one-on-one encounters with the supervisor, who served as a mentor. Notably, the frequency of these encounters was higher in Peru—a meeting every two weeks—than in Colombia—a meeting planned every six weeks, but actually happening every 7-10 weeks. In Colombia, other low-cost solutions were implemented to compensate for the less frequent in-person contact between HVs and supervisors, including bulletins with critical content to cover during the mentoring sessions and weekly reminders to HVs via text messages.

Finally, governments often encounter administrative constraints during implementation (procurement requirements, budgeting processes), and need to allow for political considerations that influence both intervention content and timing. These issues were all present to some extent in the implementation of Cuna Más.

Cuna Más staff initially had little experience with procurement processes. They were primarily focused on the production of materials and miscalculated the time and transaction costs associated with their procurement and distribution in rural localities. Consequently, the intervention was launched prior to the delivery of materials in all communities, which likely affected the quality of the home visits.

In addition, although Cuna Más was a public program, many key administrative tasks at the local level relied on community volunteer work. This led to inefficiencies and delays in payments and resource availability, and affected staff morale. Finally, political pressures led to very rapid expansion targets (often at the cost of quality) or to suboptimal resource-allocation decisions (for example, prioritizing the distribution of staff uniforms over that of manuals and toys).

To summarize, the context in which the program was implemented in Peru was very different from Colombia. In Colombia, what was implemented was an (arguably scalable) version of the original Jamaica program. In Peru, on the other hand, implementation of the program was real-life public policy—policy informed by earlier research, but public policy, nevertheless. The differences in scale were enormous: at the time when the evaluations of the two programs were conducted, the number of families covered in Peru was more than 100 times larger than that in Colombia.

What then explains the differences in effect sizes between Colombia and Peru? A variety of factors, including differences in the instruments used in the evaluation, differences in the population that received the intervention, and differences in implementation, including the training strategy, may be part of the explanation. However, the biggest difference between the evaluation samples in Colombia and Peru was arguably differences in *dosage*, or the effective number of visits received by treated children.

HVs in groups of four (the total number of HVs that there was in each town plus a replacement). In Peru, staff from the central office in Lima trained the approximately 1,000 supervisors in batches, who in turn trained their HVs (10 HVs each). Cascade training risks losses of fidelity as training moves down the cascade, particularly so if the number of people to be trained is large, as this generally leads to larger group sizes and a shorter training time.

These differences in dosage arose for two main reasons. First, program take-up among households randomly assigned to the treatment group was higher in Colombia (97%) than in Peru (66%).[12] Second, among those who received at least one visit, the number of visits received was higher in Colombia—81% of planned visits—than in Peru—50% of planned visits. In other words, effective dosage was 79% in Colombia and 32% in Peru. Therefore, it is perhaps not surprising that the effect sizes estimated in Peru are substantially lower than those in Colombia. Indeed, the ratio of effect sizes for the interventions (0.18 SDs in Colombia; 0.10 SDs for the ITT estimate in Peru) is smaller than the ratio of effective dosage, suggesting that, if anything, the effect of a given number of visits was larger in Peru than in Colombia.

In sum, as the home visiting program moved from an efficacy trial in Jamaica, to a pilot in Colombia, to an at-scale intervention in Peru, the magnitude of the impacts declined substantially. A variety of reasons probably account for this decline, and we cannot conclusively pin down the importance of different factors. Most important in our view are likely declines in quality in ways that are hard to measure and quantify. It is harder to ensure that a program is implemented exactly as intended when the total number of children who receive the intervention is 720 and the total number of home visitors is 164 (as in Colombia) than when the number of children is 64 and the number of home visitors is 3 (as in Jamaica). And it is orders of magnitude more difficult to preserve fidelity and ensure appropriate dosage when the number of children is close to 70,000 and the number of visitors is close to 7,000 (as in Peru).

4. Lessons Learned and Further Questions

This comparative analysis of the gradual scale-up of the Jamaican home visiting model from an efficacy trial to a pilot in Colombia and to a national program in Peru allows us to draw some lessons related to the scalability of this type of intervention. We focus on two types of lessons learned: (a) mitigating the threats to scalability discussed earlier and (b) research and evaluation.

4.1 Lessons Learned: Mitigating the Threats to Scalability

Lesson 1: Some Key Project Components and Attributes Are Non-Negotiable. Non-negotiable project components and attributes are the foundations of how the intervention

[12] In Colombia, the very high take-up was accomplished because the research team made a great deal of effort to ensure that all households in the treatment group received visits. In this regard, the evaluation in Colombia is somewhat similar to a lab experiment (with an intervention that could be scalable) where everyone assigned to a treatment group receives the treatment in practice. In Peru, several factors account for the lower take-up in the evaluation sample. Some of these factors are specific to the evaluation sample, others are related to take-up of the program more broadly. The sample frame for the Cuna Más evaluation was meant to exclude municipalities in which there was social conflict as well as those in which Cuna Más operated daycare services. However, because of errors in the administrative data, there were some communities in the treatment sample which did in fact have social conflict or daycare services, which only became apparent after the random assignment of municipalities and baseline data collection. Similarly, because of delays in program implementation after sample selection, some children in treatment communities were older than the age cutoff established for program enrolment (24 months). Both groups of children—those in areas where the program did not intend to operate, and those who had aged out—were kept in the evaluation sample to avoid biases, but neither received any visits. Jointly, they account for 19% of children in the treatment group of the evaluation. A further 15% of children in the treatment group did not receive visits because the program could not hire HVs in their communities or because their families turned down the program—a value that is noticeably higher than the 3% for the Colombia evaluation.

seeks to change parents' behaviors. Thus, any implementation strategy ought to retain them, even as it is adapted to different settings. We highlight two.

Training and Mentoring. HVs need adequate pre- and in-service training, as well as regular mentoring and supervision. Training should mostly be practical, rather than theoretical, and mentoring and supervision should focus on concrete things the HV does (or does not do) during the visit. The profile and background of HVs are not as relevant, so long as they have the support they need. The attitude and motivation of HVs, mentors, and supervisors are key. Many programs have a great deal of staff turnover, which is generally disruptive and reduces quality. Programs should place emphasis in retaining and minimizing turnover of effective staff.

Contents. The intervention needs to have clear, structured content that guides the HV during the visit. The lower the average education level of HVs, and the more turnover, the greater the need for structure and scaffolding to guide each visit. The program also needs to have adequate resources to develop activities and toys that are appropriate for the developmental level of each child and that promote positive interactions. The central work of the HV is through the mother/caregiver and demonstration is key during the visit. Toys need to stay in the home so that the family can continue to do the activities during the week.

Lesson 2: Political Support is Indispensable. The participation and ownership of the government is a necessary condition for at-scale implementation. There are benefits to engaging government officials early in the process to gain their buy-in and identify administrative and political constraints that are likely to arise as the program is brought to scale. While key project components should be retained, the objective should not be to design an ideal intervention but, rather, one that a government can and wants to implement at scale.

Building a real-life program involves negotiating trade-offs. One clear trade-off is between scaling up quickly and doing it with quality. The window of opportunity of political support may be narrow, and it may require scaling up at a pace that is faster than what would be desirable from a purely technical standpoint and thus, results in a sacrifice in terms of quality and fidelity. An important question is to what extent it is feasible to make corrections to a program that has been brought to scale very quickly.

Another trade-off is between a program that is narrowly focused on child development through play, and a broader program—for example, one that also includes consideration of child health and nutrition, hygiene, and intrafamily violence. Programs can become a victim of their own success: a program that was effective when it focused on child development can become ineffective when it covers too many other aspects of child and household wellbeing, even if all of these are important in their own right. It is crucial to avoid overburdening the intervention, especially taking in consideration the profile of the HVs (low technical qualifications on average), the length of the visit (one hour at best) and the time available for training.

4.2 Lessons Learned: Research and Evaluation

Lesson 1: Scalability as Part of the Research Plan A research process designed toward achieving scalability with quality would include phases of efficacy trial, quick replications, and scale-up in a way that is flexible and allows for the intervention to be fine-tuned with the findings from earlier stages. Elements from a continuous quality improvement framework (act-plan-do-study-act…) can be incorporated into this rationale—see Davis et. al (Forthcoming), Stuart (Forthcoming) and Chambers and Norton (Forthcoming) for more on designing research studies with scalability in mind.

Efficacy trials are useful to identify the non-negotiable components of an intervention and to understand the mechanisms whereby it achieves the desired outcomes. They might also allow for experimentation with relaxing certain aspects of the intervention, especially those that are more costly to implement at scale, such as the training and mentoring. Quick replicability in different contexts is necessary to move to at-scale delivery. Tracking attritors and those who do not take-up the intervention, and understanding the reasons behind these choices is also of key importance in the efficacy trial and pilot phases.

Lesson 2: Statistical Power. Studies should be designed with sufficient statistical power to be able to identify heterogenous effects along the dimensions considered relevant. Heterogeneity can be thought of not only as characteristics of the individual, but also as characteristics of the site—for example, multi-site studies and different cultural and geographic settings that might require adaptations to the intervention.

Lesson 3: Metrics. Using a common measure for key intervention outcomes (both intermediate and final outcomes) that are normed (ideally using global norms) will facilitate comparability across studies.

Lesson 4: Funders and Scientific Journals. Funders and scientific journals can play an important role aligning incentives to promote research that informs scalability by supporting (a) replication studies; (b) implementation studies; (c) non-result studies; (d) studies on metrics and measurements; and (e) medium- and long-term follow-ups that are necessary to appropriately calculate the benefits of interventions.

5. Conclusion and Outstanding Questions

Though the experiences discussed in this document shine light on the process of scaling, there are still important questions related to the examined intervention that remain unanswered and that can lower the cost of implementation at-scale. These include but are not limited to: (a) at what child age to start an intervention; (b) the effectiveness and cost-effectiveness of variations in duration, intensity, and mode of delivery (for example, group versus individual visits) and how this choice may vary by child age, geography and culture); and (c) what are the within and between treatment spillover effects (in the family, in the community). Answering these questions, in addition to applying the lessons learned from the process of scaling the Jamaica program and the theory

presented by the economic model of scaling, will help all stakeholders be better prepared to scale their own programs of interest.

References

Al-Ubaydli, O., List, J., & Suskind, D. (2019). The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments. *NBER Working Paper*, *25848*. https://doi.org/10.3386/w25848

Al-Ubaydli, O., Lee, M., List, J., & Suskind, D. (Forthcoming). The Science of Using Science: A New Framework for Understanding the Threats to Scaling Evidence-Based Policies. In J. List, D. Suskind, & L. Supplee (Eds)*, The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Andrew, A., Attanasio, O., Augsburg, B., et al. (2019). Effects of a scalable home-visiting intervention on child development in slums of urban India: evidence from a randomised controlled trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/jcpp.13171

Andrew, A., Attanasio, O., Fitzsimons, E., et al. (2018). Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia. *PLOS Medicine*, *15*(4), e1002556. https://doi.org/10.1371/journal.pmed.1002556

Araujo, M., Dormal, M., Grantham-McGregor, S., et al. (2019). *Home Visiting at Scale and Child Development*.

Attanasio, O., Cattan, S., Fitzsimons, E., et al. (2020). Estimating the Production Function for Human Capital : Results from a Randomized Controlled Trial in Colombia. *American Economic Review*, *110*(1), 48–85. https://doi.org/https://doi.org/10/1257/aer.20150183

Attanasio, O., Fernandez, C., Fitzsimons, E., et al. (2014). Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *BMJ*, *349*(sep29 5), g5785–g5785. https://doi.org/10.1136/bmj.g5785

Bitler, M., Hoynes, H., & Domina, T. (2014). Experimental Evidence on the Distributional Effects of Head Start. *NBER Working Paper*, *20434*.

Cavallera, V., Black, M., Bromley, K., et al. (2019). The Global Scale for Early Development (GSED). *Early Childhood Matters*, 80–84.

Chambers, D. & Norton, W. (Forthcoming). Sustaining Impact after Scaling Using Data and Continuous Feedback. In J. List, D. Suskind, & L. Supplee (Eds)*, The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Davis, J., Guryan, K., Hallberg, K. & Ludving, J. (Forthcoming). Studying Properties of the Population: Designing Studies that Mirror Real World Scenarios. In J. List, D. Suskind, & L. Supplee (Eds)*, The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Duncan, G., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109–132. https://doi.org/10.1257/jep.27.2.109

Gertler, P., Heckman, J., Pinto, R., et al. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, *344*(6187), 998–1001. https://doi.org/10.1126/science.1251178

Grantham-McGregor, S., Powell, C., Walker, S., et al. (1991). Nutritional supplementation,

psychosocial stimulation, and mental development of stunted children: The Jamaican study. *The Lancet*, *338*(8758), 1–5. https://doi.org/10.1016/0140-6736(91)90001-6

Grantham-McGregor, S., & Smith, J. (2016). Extending The Jamaican Early Childhood Development Intervention. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *7*(2), Article 4.

Hamadani, J., Huda, S., Khatun, F., et al. (2006). Psychosocial stimulation improves the development of undernourished children in rural Bangladesh. *The Journal of Nutrition*, *136*(10), 2645–2652. https://doi.org/136/10/2645 [pii]

Havnes, T., & Mogstad, M. (2015). Is universal child care leveling the playing field? *Journal of Public Economics*, *127*, 100–114.

Heckman, J., Liu, B. & Zhou, J. (Forthcoming). The Economics of Investing in Child Development: Applying Home Visiting at Scale. In J. List, D. Suskind, & L. Supplee (Eds)*, The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Love, J., Chazan-Cohen, R., Raikes, H., et al. (2013). What Makes a Difference: Early Head Start Evaluation Findings in a Developmental Context. *Monographs of the Society for Research in Child Development*, *78*(1).

Luo, R., Emmers, D., Warrinnier, N., et al. (2019). Using community health workers to deliver a scalable integrated parenting program in rural China: A cluster-randomized controlled trial. *Social Science and Medicine*, *239*(July), 112545. https://doi.org/10.1016/j.socscimed.2019.112545

Nahar, B., Hossain, M., Hamadani, J., et al. (2012). Effects of a community-based approach of food and psychosocial stimulation on growth and development of severely malnourished children in Bangladesh: a randomised trial. *European Journal of Clinical Nutrition*, *66*(6), 701–709. https://doi.org/10.1038/ejcn.2012.13

Nores, M., Bernal, R., & Barnett, W. (2019). Center-based care for infants and toddlers: The aeioTU randomized trial. *Economics of Education Review*, *72*(May), 30–43. https://doi.org/10.1016/j.econedurev.2019.05.004

Olds, D. (2010). The Nurse-Family Partnership: From Trials to Practice. In A. Reynolds, A. Rolnick, M Englund, & J. Temple (Eds.), *Childhood Programs and Practices in the First Decade of Life* (pp. 49–75). https://doi.org/10.1017/CBO9780511762666.004

Powell, C., Baker-Henningham, H., Walker, S., et al. (2004). Feasibility of integrating early stimulation into primary care for undernourished Jamaican children: cluster randomised controlled trial. *BMJ*, *329*(7457), 89–91. https://doi.org/10.1136/bmj.38132.503472.7C

Powell, C., & Grantham-McGregor, S. (1989). Home visiting of varying frequency and child development. *Pediatrics*, *84*(1), 157–164.

Stuart, A. (Forthcoming). Accounting for Differences in Population: Predicting Intervention Impact at Scale. In J. List, D. Suskind, & L. Supplee (Eds)*, The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

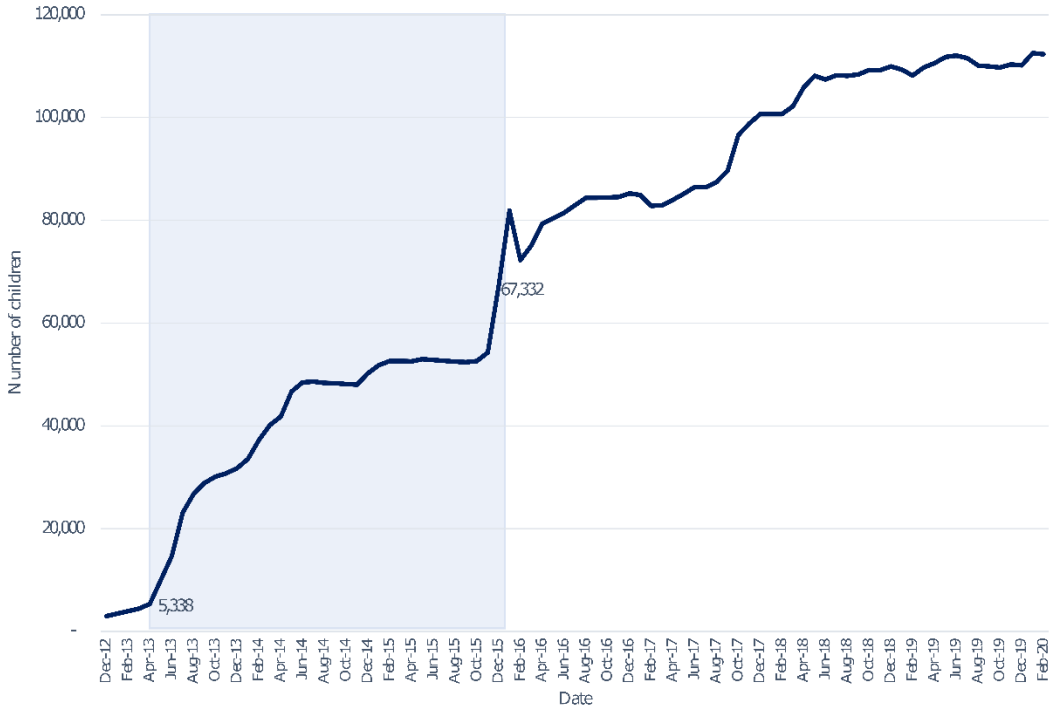
Tofail, F., Hamadani, J., Mehrin, F., et al. (2013). Psychosocial Stimulation Benefits Development in Nonanemic Children but Not in Anemic, Iron-deficient Children. *The Journal of Nutrition*, *143*(6), 885–893. https://doi.org/10.3945/jn.112.160473.Three

Walker, S., Chang, S., Smith, J., et al. (2018). The Reach Up Early Childhood Parenting

Program. Origins, Content, and Implementation. *Zero to Three Journal*, *38*(4), 37.
Walker, S., Chang, S., Powell, C., et al. (2004). Psychosocial Intervention Improves the Development of Term Low-Birth-Weight Infants. *The Journal of Nutrition*, *134*(6), 1417–1423. https://doi.org/10.1093/jn/134.6.1417
Walker, S., Chang, S., Vera-Hernández, M., et al. (2011). Early childhood stimulation benefits adult competence and reduces violent behavior. *Pediatrics*, *127*(5), 849–857. https://doi.org/10.1542/peds.2010-2231

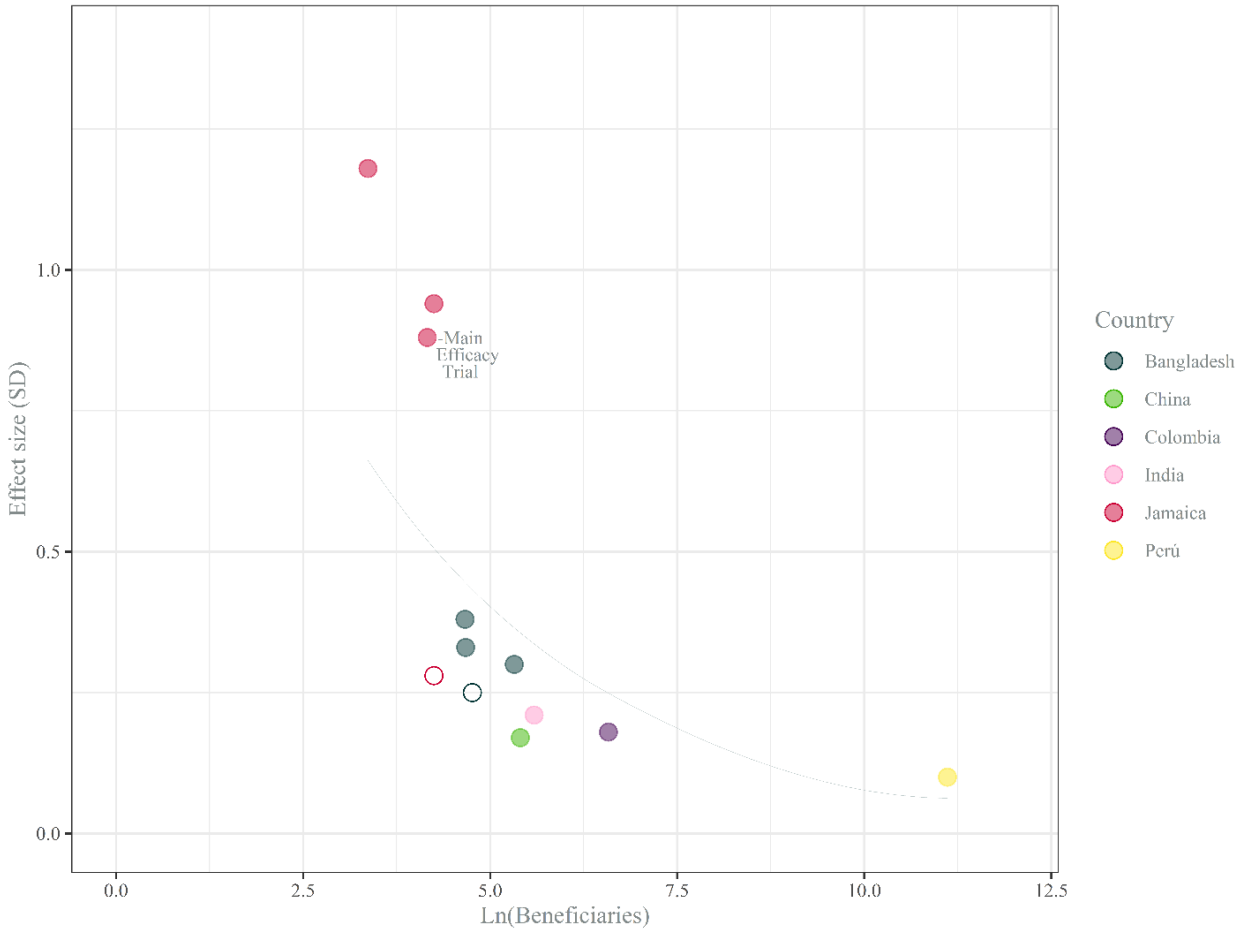
Figures and tables

Figure 1. Cuna Más Scale-Up, 2012-2020



Source: Prepared by the authors based on historic data on program coverage from the Ministry of Social Inclusion and Development – Perú, available at: http://sdv.midis.gob.pe/Infomidis/#/

Figure *2. Effect Sizes of Different Replications of the Jamaican Model*



Source: Prepared by the authors based on Table 4.

Table *1. Program Design in Jamaica, Colombia, and Peru*

	Jamaica	Colombia	Peru
Intervention	Home visits		
Intended Dosage	One hour, weekly		
Maximum Duration (months)	24	18	36
Months at Program Entry	9-24	12-24	1-24
Months at Graduation	33-48	30-42	36
HVs Training	NA	2 weeks pre-service (including practices), 1-week in-service	4 days pre-service (no practice)
Supervisors Training	DNA	6 weeks, including practices	9 days, no practice
HVs Selection	Research team	*Madres Líderes*[1] were administered a reading comprehension test by the research team and interviewed about their availability and interest	Candidates nominated by the community and selected by supervisors
HVs Characteristics			
Profile	Community health workers	63% *Madres Líderes*[1] 37% women referred by *Madres Líderes*[1]	Literacy required
Women	All	All	85%
Schooling (years)	NA	8-9	10
Area of Origin	From Kingston	From the communities	From the communities
HVs Remuneration (USD per month)	NA	~ 40 (19.5% of concurrent minimum wage)	100 during evaluation (36% of concurrent minimum wage). Increased to 115 in 2017.
Supervisors Selection	DNA	By research team	Regional open calls
Supervisors Profile			
Education	Researchers themselves	All undergraduates. Degrees in psychology or social work or, failing that, with prior experience working with families	Some tertiary education required. In practice, all had at least some post-secondary schooling
Women	All	All	~72%
Area of Origin		From Bogota	From the community or the municipality to which it belongs
Supervisors Remuneration (USD per month)	NA	~ 520-600, with 6-monthly increases between 2010 and 2011	830-900, with higher salaries paid in more remote locations
Frequency of Supervision	Weekly	Every six weeks	Biweekly
Families-to-HV	20-21	5	10
HVs-to-Supervisor	3/2	24	10

	Jamaica	Colombia	Peru
Targeting Criteria	Stunted children in urban slums in Kingston	Children 12-24 months in beneficiary families of CCT program (which targets bottom 20% of country's income distribution) in 96 semi-urban towns (2,000-42,000 inhabitants)	All children 1-24 months in rural villages (up to 2,000 inhabitants or 400 dwellings) with poverty rates of 50% or higher and stunting rates of 30% or higher
Baseline Characteristics of Program Recipients	NA		
Chronic malnutrition		14%	37%
Maternal schooling (years)		7-8	7
Homes with sewerage		90%	25%
Homes with dirt floor		6%	73%

Notes: NA: information not available. DNA: does not apply. [1] *Madres Líderes* are local leaders elected by their communities to be liaisons with the administrators of the national conditional cash transfer (CCT) program.

Table 2*. Program Implementation Status by Follow-up Evaluation**

	Jamaica	**Colombia**	**Peru**
Agency in Charge	Research team	Research team	Ministry of Social Inclusion and Development
Effective Dosage (% planned visits)	~100%	81%	50%
Effective Ratio: Families-to-HV	20-21	~5.4	9.5
Effective Ratio: HVs-to-Supervisor	3/2	~27.3	9
Effective Frequency of Supervision	Weekly	Every 7-10 weeks	Biweekly
HV Months in Program[1]	24	15 (median 17.3)	17.1
Supervisor Months in Program[1]	24	18	18.6
Total Children Served	64	720	~67,332
Number of HVs	3	144 initially; 164 by the end of pilot implementation	~10,000
Number of Supervisors	2	6	~1,000
Take-up	100%	97%	66%[2]
Costs (USD per child per year)	NA	~500	~300
Intervention Discontinued	Yes	Yes	No

Notes: NA: information not available.* Quantitative data from follow-up survey. [1] Average months in Program by follow-up survey. [2] Calculated from impact evaluation sample.

Table 3*. Program evaluation in Jamaica, Colombia, and Peru*

	Jamaica	**Colombia**	**Peru**
Research Design	RCT, child-level randomization	CRCT, town-level randomization	CRCT, municipality-level randomization
Treatment Arms	T1 = Visits T2 = Nutritional supplementation T3 = Both	T1 = Visits T2 = Micronutrient supplementation T3 = Both	T1 = Visits T2 = Visits + biweekly group meetings *T2 was not implemented, all T2 communities received T1*
Population Representativeness	Not designed to be representative of target population		
Measures	Griffiths Mental Development Scales, HOME, Height and Weight	Bayley Scales of Infant and Toddler Development III, Family Care Indicators (FCI), Height, Weight and Hemoglobin	Ages and Stages Questionnaire 3, FCI, Height and Weight
Group in Charge of Evaluation Design	Research team	Research team	Research team with inputs from technical teams at Ministry of Economics and Finance and Ministry of Social Inclusion and Development
Group in Charge of Data Collection	Research team	Specialized survey firm contracted, trained and supervised by research team	National Institute of Statistics of Peru (government agency), contracted and supervised by the Ministry of Economics and Finance. Research team participated in training and supervision.
Sample Size (Home Visiting Treatments)	129 children (32 in T1 + 32 in T3) 127 children analyzed (30 in T1 + 32 in T3)	1440 children (360 in T1 + 360 in T3) 1263 children analyzed (318 in T1 + 319 T3)	5,339 children (3,530 in T1 + T2) 4,685 children analyzed (3,192 in T1 + T2)
Attrition	0%	9.4% in Control 11.7% in T1 13.9% in T2 11.4% in T3	8.6% in Control 9.3% in T1 + T2
Impact (in Standard Deviations, SD)	<u>Intent-to-treat</u> 0.88 SD in Developmental Quotient (DQ)[1]	<u>Intent-to-treat</u> (T1) 0.26 SD in cognition 0.22 SD in receptive language 0.18 SD in aggregate measure (cognition, receptive language, expressive language, fine motor development) T3: not statistically significant from T1.	<u>Intent-to-treat</u> (T1 + T2) 0.08 SD in cognition 0.10 SD in language 0.10 SD in total score[2] <u>Treatment-on-the-treated</u> (T1 + T2) 0.15 SD in cognition 0.15 SD in language 0.14 SD in total score[2]

Notes [1]The DQ score includes the locomotor (gross motor), hand and eye coordination (fine motor), hearing and speech (language), and performance (cognition) subscales. [2] Total score is an aggregate measure of cognition (problem solving), language, fine motor, gross motor and personal-social development.

Table 4*. Replications of the Jamaican Model and Effect Sizes, in Standard Deviations (SD)*

Country	Reference	n (T)	Outcome	Impact
Jamaica	Powell & Grantham-McGregor, (1989)	58 (29)	Griffiths Mental Development Scales, first edition; Developmental Quotient (DQ)	1.18 SD
	Grantham-McGregor et al., (1991)	127 (30)		0.88 SD
	Walker et al., (2004)	130 (63)		0.28 SD (not significant)
	Powell et al., (2004)	129 (65)		0.94 SD
Bangladesh	Hamadani et al., (2006)	193 (92)	Bayley Scales of Infant Development, second edition; Mental Development Index (MDI)	0.33 SD
	Nahar et al., (2012)	322 (59)		0.30 SD
	Tofail et al., (2013)	non-anemic children: 196 (104)		0.38 SD
		iron-deficient (anemic) children: 216 (110)		0.25 SD (not significant)
Colombia	Attanasio et al., (2014)	1,263 (318)	Bayley Scales of Infant and Toddler Development, third edition; aggregate measure of cognition, language and fine motor development	0.18 SD
Peru	Araujo et al., (2019)	4,685 (3,192)	Ages and Stages Questionnaire, third edition; total score of cognition, language, fine and gross motor, and personal-social.	0.10 SD
India	Andrew et al., (2019)	378 (191)	Bayley Scales of Infant and Toddler Development, third edition; aggregate measure of cognition, receptive and expressive language, and fine motor development	0.21 SD
China	Luo et al., (2019)	390 (190)	Bayley Scales of Infant and Toddler Development, third edition; aggregate measure of cognition, receptive and expressive language, fine and gross motor, and socio-emotional development	0.17 SD

Note: Table based on Table 1 in Grantham-McGregor and Smith (2016). It adds studies published since and excludes those without randomized evaluations or where home visits were combined with another intervention. We report effect sizes on aggregate developmental scores, irrespective of whether there are impacts on its subscales. In all cases, we report sample and effect sizes for the visits only treatment group (compared to the control). We convert impact sizes into SD using the SD of the control group at follow-up. Effect sizes adjusted for baseline covariates reported in all cases. Specific details on the studies reported are available from the authors upon request.